The C.A.Ve. Software Tool for Genome Assembly.

Dhileepan Sivam, Peter Myler, PhD.

Division of Biomedical & Health Informatics, University of Washington, Seattle WA.

SUMMARY:

Misassembly due to the presence of repetitive DNA sequences often complicates the contig assembly stage of genome sequencing. Accurate physical representations of chromosomes, such as restriction maps or contig maps, provide a means for validating the sequence assembly and clarifying alignment ambiguity. The Contig Assembly Verifier (CAVe) software tool allows the researcher to automatically reconcile a sequence assembly with a physical representation, and, if needed, to modify the assembly using an intuitive graphical user interface.

INTRODUCTION:

Whole chromosome shotgun assembly is highly precise on the individual nucleotide level; ten-fold coverage can yield approximately 100 percent accuracy at each base pair. However, repetitive sequence can lead to misassembly when constructing sequence contigs, resulting in an inaccurate whole chromosome sequence. Misassembly is particularly common when attempting to assemble sequences containing long (> 500 base pair) repeats. For this reason, many current assembly methodologies are of questionable stand-alone accuracy.

Optical mapping, one example of a physical representation, is a single molecule approach for the construction of ordered restriction maps. Optical maps provide a useful template for the assembly of sequencing data, especially in DNA regions containing long repeat sequences. The restriction fragments established by the optical map can serve as landmarks by which to validate the ordering of the restriction fragments determined from the contig assemblies.

Discrepancies between the ordering of the physical map fragments and assembly fragments are resolved by attempting to reorder the assembly fragments into a pattern that better matches the pattern observed in the physical map.

PROBLEM:

Current tools require the manual verification, and reordering, of the sequence assembly against the physical representation. As the sequencing data continually changes, the verification and reordering must be constantly refined; such continual manipulation can prove extremely time-consuming and difficult.

ALGORITHM:

The Contig Assembly Verifier tool automatically verifies the positioning of the restriction fragments from the assembled sequence (**ASSEMBLY FRAGMENTS**) against the fragments from the physical representation (**PHYSICAL FRAGMENTS.)** The verification is performed by a scoring system that enumerates and scores, based on size and ordering, various pairings of assembly fragments to physical fragments.

Initially, the assembly fragments are coupled to the physical fragments in a pairwise manner using length as the scoring criterion, with a cutoff score excluding improbable pairings. The pairwise matches, known as **SEED MATCHES**, are then ranked in descending score order.

Each pairwise match is used in a **SEED SCORING** run to determine the highest-scoring physical fragment pairing for each assembly fragment. The Seed Matches, which are based on size alone, cannot predict the best pairing, because many fragments may have the same approximate size. The Seed Scoring algorithm assigns higher weights to placements that are flanked by other high-scoring placements, thus taking into account fragment ordering.

The weight assigned to each pairing of an assembly fragment with a physical fragment is preserved and incremented through each Seed Scoring run. When an assembly fragment is matched with a physical fragment in the Seed Pairing, the pairing is given a positive weight, based purely on size similarity. If the assembly fragments adjacent to the Seed Pairing assembly fragment are seeded with physical fragments adjacent to the Seed Pairing physical fragment, the pairing will acquire an increased weight during the Seed Scoring runs based on the similarity in adjacent fragments.

The highest weight pairings will determine the positioning of the assembly fragments on the physical map. If misassembly has occurred, assembly fragments may be re-ordered to fit the physical map.

CONCLUSION:

The Contig Assembly Verifier tool validates sequence assembly and provides alternate assembly suggestions based on correlation of sequencing data to a physical chromosome representation.